DOI：10.14005/j.cnki.issn1672-7673.20230320.002

用于 McIntosh 分类的太阳黑子数据收集与实验验证

周美林[1,2,3]，钟立波[2,3*]
（1. 中国科学院大学，北京 100049；2. 中国科学院光电技术研究所，四川 成都 610209；
3. 中国科学院自适应光学重点实验室，四川 成都 610209）

摘要：作为预测太阳活动的重要依据，太阳黑子的麦金托什(McIntosh)分类由于其中某些类别与耀斑爆发有着紧密联系而应用广泛。随着数据量的快速增加，自动化进行太阳黑子的麦金托什分类已成为迫切需求。使用太阳动力学观测站(Solar Dynamics Observatory，SDO)上的日震与磁场成像仪(Helioseismic and Magnetic Imager，HMI)提供的 720s-SHARP(Spaceweather HMI Active Region Patch，SHARP)系列数据产品和美国国家海洋和大气管理局(National Oceanic and Atmospheric Administration，NOAA)提供的太阳区域摘要(The Solar Region Summary，SRS)信息作为用于麦金托什分类的图像数据来源和标签数据来源，首先在仅有 7 年数据 Sharp 数据库基础上进行扩充，建立一个完整太阳周期(时间跨度为 12 年)且经过数据清洗的有效太阳黑子 newSharp 数据库；其次根据太阳黑子图像的特征，采取一系列如按活动区分配数据等预处理操作，并证明其科学性和必要性；最终使用卷积神经网络(Convolutional Neural Network，CNN)中 4 种经典的分类神经网络模型将 Sharp 和 newSharp 进行麦金托什(McIntosh)分类对比实验。实验结果表明，newSharp 相比于 Sharp，除了数据量有显著提高，同时有效样本的加入和无效样本的清洗使得大部分类别的加权 F_1 分数有所提升，少类的加权 F_1 分数实现 0 的突破；其中 McIntosh-p 的加权 F_1 分数整体提升最大，验证了建立完整可靠的数据库和使用科学合理的实验方法的有效性，能较好实现自动化且端到端地处理实际观测到太阳黑子图像的麦金托什分类任务。

关键词：太阳黑子；McIntosh 分类；卷积神经网络；Sharp 数据集
中图分类号：TP391.4 **文献标识码**：A **文章编号**：1672-7673(2023)04-0353-16

太阳活动中耀斑爆发可以引起空间环境变化，对人类活动产生巨大的影响[1-6]，而国际上公认且应用广泛的麦金托什分类中较为复杂的太阳黑子类别 D，E，F 与耀斑爆发有着紧密联系[7]。因此，太阳黑子的麦金托什分类可以作为预测太阳耀斑的重要依据；同时由于快速增长的数据量[8-9]，如何高效对太阳黑子自动进行麦金托什分类已成为太阳物理领域的迫切需求。

太阳黑子麦金托什分类主要经历了从专家手动、图像处理到深度学习方法。以往麦金托什自动分类方法[10-13]由于主要使用全日面图像将分类任务分成太阳黑子识别、聚类/分组、分类等任务分步进行，其中分类任务主要根据 McIntosh-Zpc 分类规则作为决策树的决策标准完成，但前期需要大量专家提取数据特征，实际操作复杂且精度较低。2008 年，文[13]建立决策树进行 McIntosh-Z/c 分类，用简单的全连接神经网络模型完成 McIntosh-p 分类，前期同样需要大量专家手动依次提取图像中黑子群的特征并以此作为决策树或神经网络的输入，最终结果往往受聚类算法和类别不均衡影响，导致部分类别精度几乎为 0；此外，由于使用数据的时间跨度远远短于 11 年太阳黑子周期(太阳黑子的活动变

基金项目：国家自然科学基金（11727805）；中国科学院青年创新促进会（2022386）；中国科学院光电技术研究所前沿研究基金（C21K002）资助.
收稿日期：2022-12-21；修订日期：2023-01-17
作者简介：周美林，男，硕士，主要研究太阳黑子图像处理. Email：1321334479@qq.com.
* 通信作者：钟立波，女，副研究员，主要研究太阳活动区自适应光学图像事后处理技术. Email：zhonglibo@ioe.ac.cn.

化规律具有 11 年的周期性，本文将其作为可以包含一个周期内太阳黑子特征的时间跨度参考值)，包含的黑子种类和数据较少，无法涵盖足够的黑子特征，模型可以利用的样本较少，最终分类结果缺乏可信度。2000 年以来，深度学习典型算法中的卷积神经网络可以从经过简单预处理的数据甚至是原始数据中，学习到本质的、抽象的和高阶的特征，并成功应用于图像中目标和区域的检测、分割和识别任务[14-17]，因此一直受到广泛关注。2019 年，文[1]基于 SDO/HMI 的连续光谱全日面图，用目标检测的方法进行麦金托什分类，通过大量人工手动标注 2013~2016 年的太阳黑子图像，获得 8 800 个标签数据，最终仅通过 2017 年共 431 个黑子进行测试，部分类别数量甚至为 0，同时由于太阳黑子图像的连续性，随机分配数据集往往造成分类精度虚高。2020 年，文[6]基于大气成像组件(Atmospheric Imaging Assembly，AIA)全日面裁剪图像，使用 ResNet-50 对获得的 550 张样本进行磁分类，结果表明，尽管训练精度可达 97%，测试精度仅有 30%，随机对 26 个黑子进行测试，由于类别数量之间最大相差超过 2 倍的类别不均衡现象，分类结果两极分化严重。可以看出，深度学习算法具有较强的数据依赖性，对于太阳黑子麦金托什分类任务而言，以往工作中出现的问题主要来源于数据量少、数据集划分方式不合理等。总而言之，目前的分类算法采用的数据来源众多、缺少统一标准、类别数多(60 类)而数据量少、类别不均衡等主要因素导致解决方案复杂和模型过拟合严重；另一方面，对于太阳黑子数据的分配方式、评价标准、数据预处理等方法合理性的问题也层出不穷，在两者共同作用下，分类结果不理想，因此往往难以通过深度学习实现精确且自动化的麦金托什分类。

综上所述，来自数据和方法方面的问题是目前实现自动分类目标的首要挑战。本文根据以往工作的问题和难点，以及实验流程中的科学性操作，分别从数据和方法方面解决问题：首先重点解决标准数据库的问题，使用局部日面图像建立完整太阳周期(时间跨度 12 年数据集)，且经过数据清洗，同时保留一定现实数据特征的太阳黑子 newSharp 数据库；另一方面，结合太阳黑子数据特点，对样本进行 0-padding 和视场统一化等预处理，再使用活动区(Active Region，AR)编号进行科学合理分配数据，避免以往工作中因随机分配方式出现的数据集交叉污染情况，并采用基于类别数量的加权 F_1 分数作为评价指标，既避免以往仅使用分类准确率(Accuracy)而未同时关注查准率(Precision)与查全率(Recall)，也避免了以往使用平均准确率使得数量极少类别贡献不合理、不具备普遍性与说服力的表现影响分类结果。最终本文选取并使用卷积神经网络中一系列经典的分类神经网络模型进行太阳黑子麦金托什自动化分类实验，以充分验证 newSharp 数据库和实验操作的有效性和必要性，为未来实现基于实际复杂数据集且端到端的太阳黑子麦金托什分类任务打下基础。

1 麦金托什分类标准

目前，国际上公认的太阳黑子群分类有三大标准，分别是威尔逊(Wilson)山磁分类[18]、苏黎世(Zurich)分类[19-20]、麦金托什分类[21]。

具体而言，威尔逊山磁分类主要基于磁场极性将太阳黑子分为 α，β，γ，β-γ，δ，β-δ，β-γ-δ 和 γ-δ 等 8 类。相较之下，苏黎世分类更关注太阳黑子的演化顺序与形态特征并将其细分为 A，B，C，D，E，F，G，H 和 J 等 9 类。观察发现，即使是最活跃的 F 类，爆发大耀斑的概率依然很低[22]。麦金托什分类在修正的苏黎世分类(即 A，B，C，D，E，F 和 H，共 7 类)基础上，额外引入更能细分太阳黑子且关联耀斑爆发的参数：描述太阳黑子组内最大黑子形态的 p 参量(有 x，r，s，a，h 和 k，共 6 类)和描述太阳黑子组内部紧密程度的 c 参量(有 x，o，i 和 c，共 4 类)，如图 1，三者共同组成麦金托什的 Zpc 分类规则。研究表明，麦金托什分类中的 Dkc，Eki，Ekc，Fki 和 Fkc 类别与 m.x 级 X 射线事件的爆发率联系极高[23]，所以可以通过太阳黑子分类预测耀斑爆发等剧烈太阳活动，且这种方法对黑子群的形态特征描述最为全面，对人类观测理解太阳活动与极端空间天气的预警有重要作用。因此，麦金托什分类是目前在天文学中应用最多，也是太阳物理学家在黑子群分类中使用最广的方法[1]。

基于深度学习方法进行太阳黑子麦金托什分类的整体流程是首先建立足够多数量和类别的有效数据库，至少包含一个太阳周期的黑子特征，每一个数据样本由包含活动区的局部光球层图像及对应的麦金托什分类标签组成；将准备好的数据库样本进行科学合理的训练集-验证集-测试集划分以及预处理；然后输入神经网络模型并获得训练结果，由完整可靠的数据库训练充分的网络模型可以实现自动化太阳黑子麦金托什分类任务。可以看出，该过程的首要任务在于获得完整有效的太阳黑子数据库。鉴于以往工作中出现的数据量少、数据集划分方式不合理的问题，本文通过数据扩充、数据预处理、数据合理划分等步骤致力于构建更完善可靠的数据库，为后续太阳黑子麦金托什分类任务的实现奠定基础。

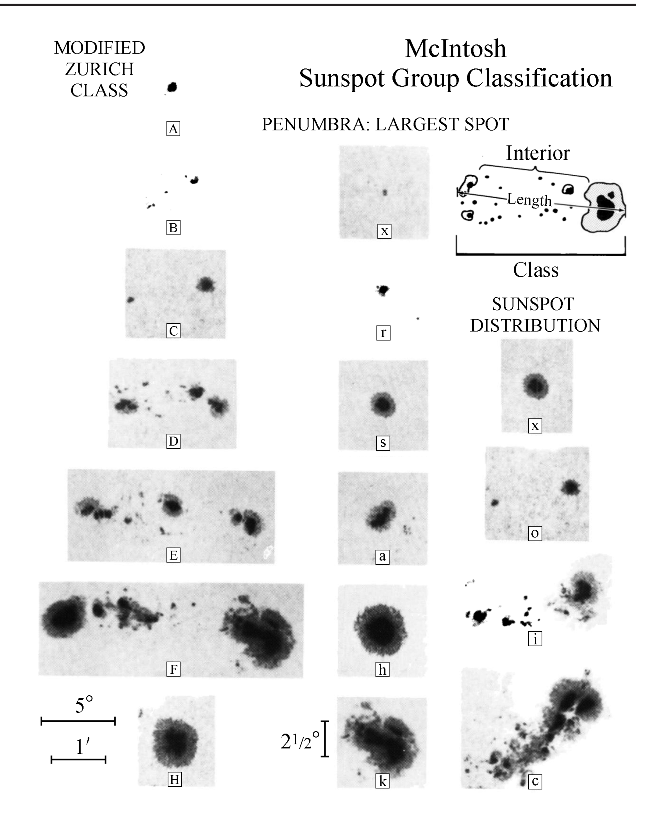


图 1　McIntosh-Zpc 分类规则[21]

Fig. 1　Classification code of McIntosh-Zpc[21]

2　数据准备与预处理

2.1　太阳黑子数据库 Sharp

与以往工作中使用全日面图像不同，本文使用文[24]整理的局部日面图像数据，由空间环境人工智能预警创新工坊提供，同时于 2021 年 6 月 21 日公布作为阿里云天池大赛的太阳黑子群磁分类竞赛的官方数据集。通过对 2010～2017 年共 15 641 个太阳黑子 FITS(Flexible Image Transport System)文件解压与匹配对应的 McIntosh 标签，我们获得了可用于麦金托什分类的原始太阳黑子数据库，并将其命名为 Sharp。

Sharp 数据库的图像数据来源是由搭载在太阳动力学观测站上的日震与磁场成像仪提供的空间天气 HMI 活动区域数据产品(Spaceweather HMI Active Region Patch，SHARP)，旨在通过收集、存储、跟踪和分析局部日面活动区图像以研究太阳活动的变化情况[25-26](作为区分，Sharp 代指原始麦金托什太阳黑子数据库，SHARP/720s-SHARP 代指空间天气 HMI 活动区域数据产品)。SHARP 系列数据产品包括时间间隔为 12 min 的磁图和可见光图像，提供了活动区域地图，同时包含整个生命周期的自动跟踪磁场强度[27]，并存储为 FITS 格式文件，可由两个主键索引：时间(T_REC)和 HMI 活动区编号(HARPNUM)。FITS 格式是一种定义和编码数据的方法，1982 年由国际天文学联合会(International Astronomical Union，IAU)确立，以便于世界各天文台之间的天文图像数据传输和交换[28]。与太阳黑子麦金托什分类相关的关键字参数如表 1。

表 1　FITS 文件关键字参数

Table 1　Keywords description of FITS file

Parameter name	Description
SIMPLE	Whether file conforms to FITS standard
DATE_OBS	Observation date
NASIS1/2	Width/Height of original image
HARPNUM	HMI active region patch number
NOAA_AR	Best single matching NOAA AR number
NOAA_ARS	Comma-separated list of all matching NOAA AR numbers
NOAA_NUM	Number of entries in NOAA_ARS

Sharp 数据库的标签信息来自美国国家海洋和大气管理局，美国国家海洋和大气管理局每天将太阳黑子群麦金托什分类信息以 SRS 文件的形式实时发布在 http://www.solarmonitor.com 上，并且由于该网站发布信息的及时性和完整性，得到了大部分天文研究机构的关注和认可。美国空间天气预报中心(Space Weather Prediction Center，

SWPC)汇编的 SRS 是美国国家海洋和大气管理局和美国空军(United States Air Force，USAF)的联合产品，每天 0030 UTC 时发布，提供前一天在日面上观测到的活动区域的详细说明。SRS 由美国空间天气预报中心在分析和整理美国空军太阳光学观测网(Solar Optical Observing Network，SOON)的所有单独报告后汇编。SRS 文件关键字参数如表 2。

表 2 SRS 文件关键字参数

Table 2 Keywords description of SRS file

Parameter name	Description
Nmbr	NOAA active region number
Location	Sunspot group location
Lo	Carrington longitude of the group
Area	Total corrected area of the group
Z	McIntosh classification of the group
LL	Longitudinal extent of the group in heliographic degrees
NN	Total number of visible sunspots in the group
Mag Type	Magnetic classification of the group

本文通过将太阳动力学观测站日震与磁场成像仪的 FITS 文件和美国国家海洋与大气管理局 SRS 的 McIntosh 标签信息两者进行匹配，具体操作过程为(1)从 NOAA 以 FTP 的方式获取 2010~2017 年所有 SRS 单独文件；(2)将获取的 SRS 文件信息过滤提取关键字数据，并按年份进行汇总，最后输出 csv 文件；(3)遍历 FITS 文件数据并解压为 JPG 格式，以日期和 NOAA_AR 搜索对应日期 SRS 文件相同活动区编号的麦金托什信息并命名。最后整理得到的 Sharp 数据库中共有 15 641 个可用样本，包括 54 类麦金托什太阳黑子数据，图 2 是 Sharp 数据库中 2015 年 6 月 22 日 0 点且 HARP 编号为 5692、活动区为 12371、麦金托什分类为 Fkc 的图像示例。

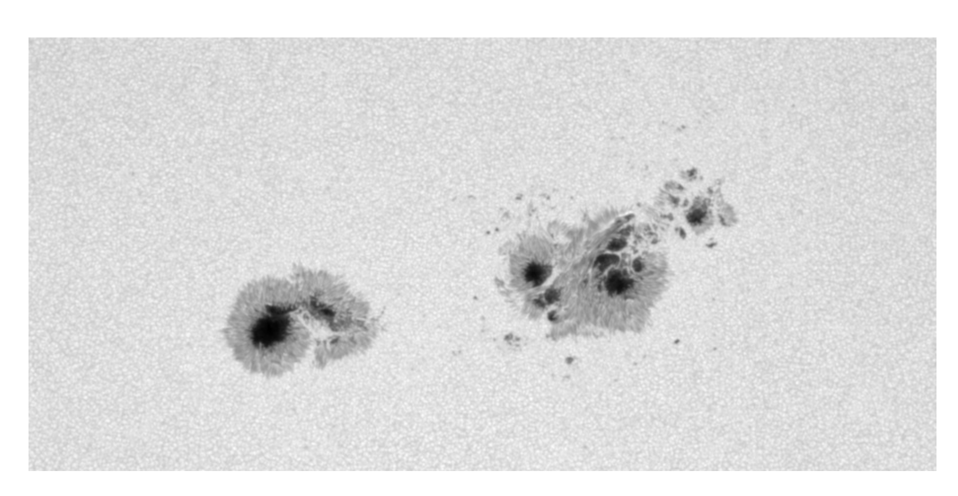


图 2 Sharp 数据库中来自 2015 年 6 月 22 日 0 点，HARP 编号为 5692、活动区为 12371、麦金托什分类为 Fkc 的图像示例

Fig. 2 Image example from Sharp dataset with date of 2015. 06. 22. 00:00:00，HARPNUM of 5692，NOAA of 12371 and McIntosh class of Fkc

然而，上述用于麦金托什分类的原始 Sharp 数据库依然存在一系列问题，无法满足太阳黑子分类任务的要求。首先，Sharp 数据库的数据时间范围仅有 7 年，远远小于一个太阳黑子周期，使得 Sharp 数据库包含的黑子特征不够完整，同时数据量较少，可信度较低；其次，Sharp 数据库存在麦金托什分类类别数据不均衡的问题，同时存在一系列无效数据需要过滤；此外，通过数据来源得知，Sharp 数据库具有较大的扩充可能性。因此，本文在 Sharp 数据库基础上进一步扩充及清洗数据。

2.2 太阳黑子数据库 newSharp

2.2.1 数据扩充

与 Sharp 数据库类似，扩充数据同样使用来自 SDO/HMI 提供的 720s-SHARP 系列数据产品(hmi.sharp_720s-Space Weather HMI Active Region Patch)[27]，所有数据文件从网站 http://jsoc.stanford.edu/ 下载，均采用 FITS 格式，并将扩充数据库命名为 newSharp。数据选择满足以下标准：(1)时间范围为 2010 年 5 月至 2022 年 8 月；(2)图像数据每 96 min 拍摄一次；(3)只有当一个 SHARP 编号同时对应

于一个 NOAA AR 时，才会选择 SHARP 数据。观察发现，原 Sharp 数据库的数据实际时间范围是 2010 年 5 月至 2017 年 12 月，同时在此时间范围内存在数据缺失，即大量可用数据被忽略，故本文将数据扩增过程分为两大步：2010~2017 年获取遗漏的数据，2018~2022 年获取每天数据。两者具体下载流程如下：(1)从 2010 年 5 月 1 日开始下载当天 0 点的 FITS 数据，并查看是否包含在原 Sharp 数据库中，如有则删除，进行下一项数据，若无则进一步获取 FITS 文件中“NOAA_NUM”关键字(代表该图像块中包含黑子所在的活动区个数)，若大于 1 则删除，进行下一项，若等于 1 则以 96 min 的间隔获取该活动区当天所有 FITS 数据，进行下一项数据；(2)从 2018 年 1 月 1 日开始，无需判断是否包含在原 Sharp 数据库中，直接下载当天 0 点 FITS 文件并解析其“NOAA_NUM”数值大小，若大于 1 则删除，进行下一项，若等于 1 则同样以 96 min 的间隔获取该活动区当天所有 FITS 数据，进行下一项数据。

完成 newSharp FITS 数据和 SRS 数据扩充后，将 FITS 文件解压，获取其“NOAA_AR”和“HARPNUM”参数，并根据“NOAA_AR”和时间作为连接与 SRS 文件信息对应获得该活动区内黑子群的 McIntosh 编号。最终，从 15 641 张太阳黑子图像的 Sharp 数据库扩增到 107 153 张太阳黑子图像的原始 newSharp 数据库。至此，一个完整太阳周期(2010 年 5 月 1 日至 2022 年 8 月 8 日)的太阳黑子数据库 newSharp 初步建立完成。然而，其中夹杂着大量无法使用的数据需要进一步处理。

2.2.2　数据清洗

通过观察，newSharp 中混入大量无法使用的图像数据，需要进一步清洗与过滤。首先是无效数据，如图 3(a)，由于设备等因素影响形成黑图，数据库无法直接使用，需要删除；如图 3(c)，图像中混入了其他组的黑子(包括但不限于突然出现、从边缘进入、逐渐平移进入等)，也需要人工删除；此外，newSharp 中也存在大量处于极端的日面边缘且黑子不明显(甚至不存在黑子)的图像数据,如图 3(b)，本文考虑到黑子群具有一定长度，利用对应的 FITS 文件获取“LON_MIN”和“ LON_MAX”经度关键字信息来过滤此类数据，具体规则是筛选并删除 newSharp 中经度大于 80°(黑子不明显)或者经度大于 75°且经度范围小于 15°(黑子处于边缘且畸变严重)的太阳黑子图像数据。

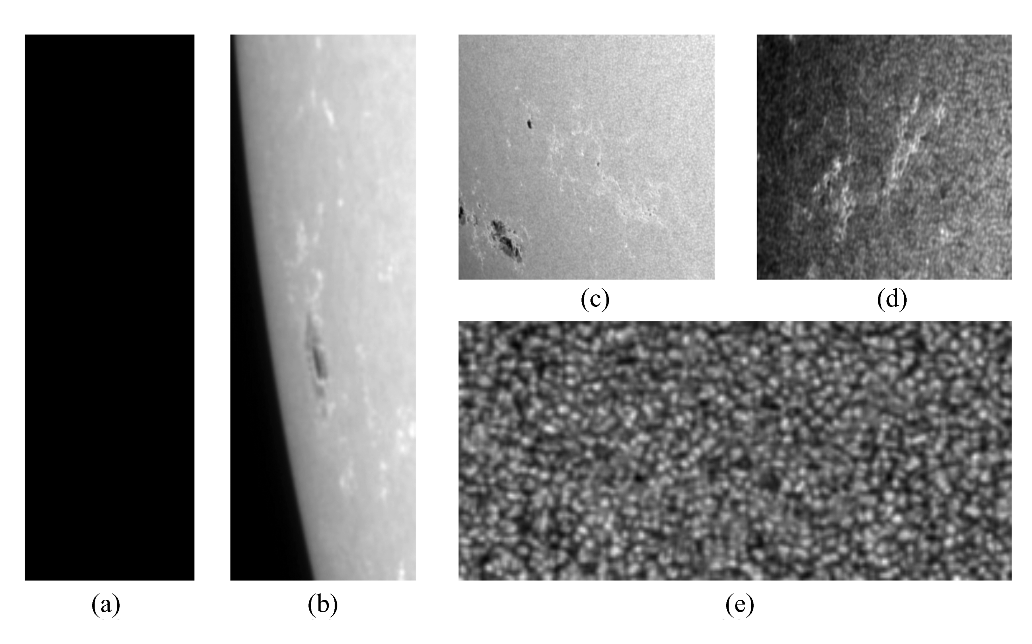


图 3　原始 newSharp 中的被清洗数据示例。(a)黑图；(b)黑子不明显图；(c)混入其他黑子图；(d)受临边昏暗影响图；(e)受光球层噪声影响图

Fig. 3　Image examples of cleaned data from original newSharp. (a) Black image; (b) image with inconspicuous sunspots; (c) image with other sunspots; (d) image affected by limb darkening; (e) image affected by photosphere noise

过滤无效数据之后，我们继续对污染严重的数据进行清洗，例如：图像模糊、来自临边昏暗或

光球层噪声的影响突然达到不可忽略的程度(该程度由主观决定，如图 3(d)、图 3(e))等图像数据，需要人工进行考量并删除。总而言之，newSharp 集合了太阳黑子在球体表面位移、旋转，同时不断演化，以及一定日面边缘位置与噪声等一系列因素在内，反应了真实的观测结果，具有普适性和合理性。

数据库中的每个数据由图像和标签组成，标签的准确性直接影响网络的训练结果，因此，我们对标签的有效性进行复核。由于获取的图像是每天间隔 96 min 的所有数据，而标签是与当天 0030 UTC 发布的 SRS 标签数据相匹配，故每天仅有固定时刻 SRS 提供的一个麦金托什标签，因此在将 0030 UTC 标签信息同样赋予其他时刻的图像数据时，没有考虑到此时黑子极有可能早已演化为另一类型。即由于太阳黑子演化的连续性，在赋予下一个 SRS 标签之前，黑子已经演化为另外的类型而导致标签信息有误。如图 4，选取 3 张图像(20141102_Hrx_4751_12200_093600.jpg，20141102_Hrx_4751_12200_222400.jpg 和 20141103_Axx_4751_12200_000000.jpg)作为示例，可以发现，同一天内相同活动区、不同时刻的黑子形态已经不同，然而标签信息却使用前一个 SRS 提供且针对 0030 UTC 时刻的太阳黑子图像标签数据，这显然存在错误。

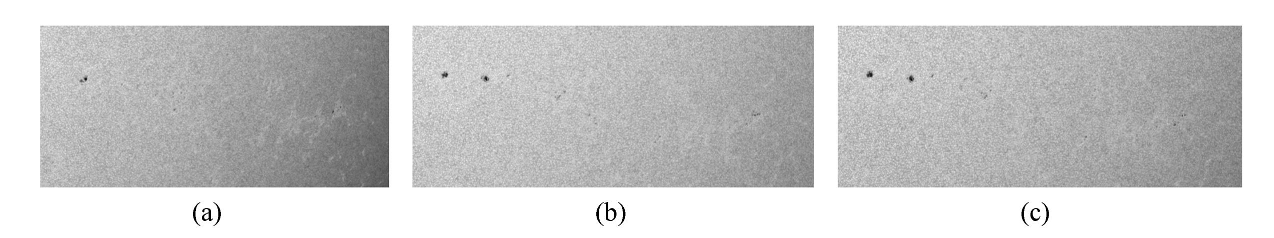


图 4 原始 newSharp 中的 McIntosh 标签信息有误数据示例。(a)20141102_Hrx_4751_12200_093600.jpg；(b)20141102_Hrx_4751_12200_222400.jpg；(c)20141103_Axx_4751_12200_000000.jpg

Fig. 4 Image examples of data with wrong McIntosh label from original newSharp. (a) 20141102_Hrx_4751_12200_093600.jpg; (b) 20141102_Hrx_4751_12200_222400.jpg; (c) 20141103_Axx_4751_12200_000000.jpg

针对上述情况，我们经过一系列观察发现共同特点，汇总起来具体可以分为两种情况：(1)开始分类正确，由于黑子进化导致中间过程中黑子分类标签错误，即活动区相同而前后分类标签不同；(2)开始便分类错误，即在一个 SRS 更新周期内，活动区相同，前后分类标签不同。根据此信息可以编写程序将前后两天的标签数据进行比较，若前后两天标签数据未改变则进行下一项；若前后标签信息不同，则表明处于上述两种情况之一。由于需要太阳物理专家的专业指导，而目前暂未有较为权威的过程中间太阳黑子麦金托什类别标签数据，故本文仅保留 0 点数据而移除中间过程中标签存疑或有误的数据，最终以此规则在 newSharp 数据库中进行迭代。

完成上述操作后，基本保证 SRS 标签信息匹配了对应的太阳黑子图像数据。至此，一个完整太阳周期的有效太阳黑子数据库 newSharp 建立完成。值得强调的是，考虑到需要足够数据量、面向实际任务情况等，newSharp 中包含一定程度的边缘黑子和受临边昏暗影响的图像。

综上所述，本文从原始 15 641 张太阳黑子图像的 Sharp 数据库扩增到共 107 153 张太阳黑子图像的原始 newSharp 数据库，但由于发现 SRS 信息未匹配的 15 847 条数据，以及活动区编号不匹配的 54 条数据，移除后经人工清洗和过滤黑图及边缘黑子不明显图 19 446 张，后又发现 2 475 组(每组约 15 张图片)前后标签有差异的图片，选择保留当天 0 点数据，其余数据移除，最终建立共 40 246 张有效样本的太阳黑子 newSharp 数据库。

2.3 太阳黑子数据预处理

数据预处理方面，我们观察发现 newSharp 中存在大量处于日面边缘但依然有效的数据(如图 5(a))，考虑到这些太阳黑子图像特点，除了常规的数据增强以外，本文通过选择 0-padding 的方式进一步消除宇宙背景的影响，如图 5(b)，并以训练好的 ResNet-18 神经网络模型为例，对较为靠后的第三模块输出特征图进行观察，靠后的模块提取的特征往往对于最终分类结果较为重要，如图 5(c)，

ResNet-18 能较好地提取包括相应太阳黑子所在区域特征在内的一系列特征，并以此作为分类结果的重要依据。

此外，由于神经网络模型的输入有特定尺寸要求(例如 ResNet 网络的输入尺寸为 224×224)，直接将太阳黑子数据输入神经网络会导致太阳黑子形态和实际尺寸发生改变。但实际上，黑子相对大小对于麦金托什分类而言较为重要，故本文进一步将太阳黑子数据的视场进行统一操作，还原其真正的视场大小。首先获取 newSharp 数据对应 FITS 文件中的“CDELT1/2”分辨率参数，结果显示其数值均为 0.504，可知 newSharp 中的图像数据处于同一视场大小，不能直接进行简单的放缩操作。其次分别获取 newSharp 中图像数据的最大长度和最大宽度，将两者进行对比继续选择最大的尺寸，以此作为特定网络模型输入尺寸的最大参考值，供其余图像在该尺寸下按原比例进行缩放，例如拥有最大宽度的图像数据在 224×224 中将 224 作为宽度值，高度按原比例缩小，同理其余图像首先将较大的尺寸在 224 × 224 中按照与最大宽度的比率缩小，而后宽度按照原比例缩小，如此可以保证在 224 × 224 的尺寸下，输入图像中黑子形态与视场大小保持一致。最后在保证原视场大小的基础上对图像数据进行 0-padding 操作，如图 6，即填充(padding)后的图像尺寸应为特定网络模型输入尺寸。

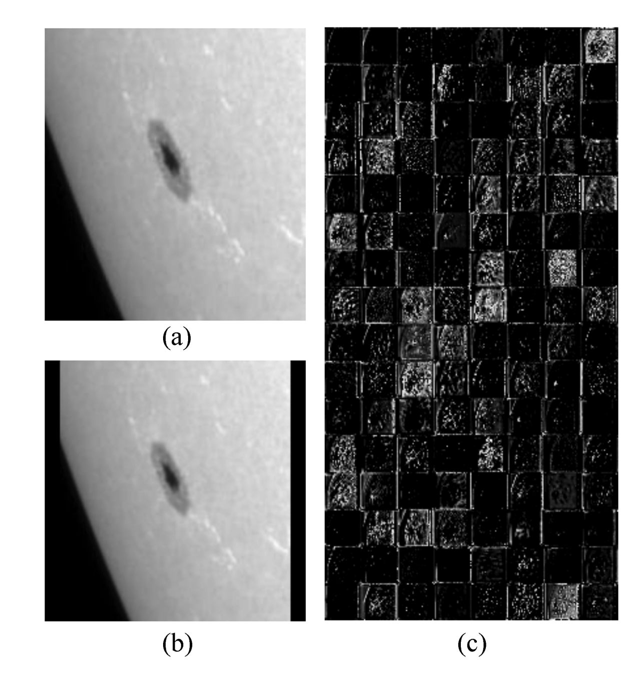


图 5　0-padding 处理后的太阳黑子图像数据以及 ResNet-18 特征提取图例。(a)原始太阳黑子图像；(b)经过 0-padding 后的太阳黑子图像；(c)ResNet-18 第三模块特征图

Fig. 5　Image examples of sunspot data after 0-padding and feature extracting result of ResNet-18. (a) Original sunspot image; (b) sunspot image after 0-padding; (c) feature map of ResNet-18′s third block

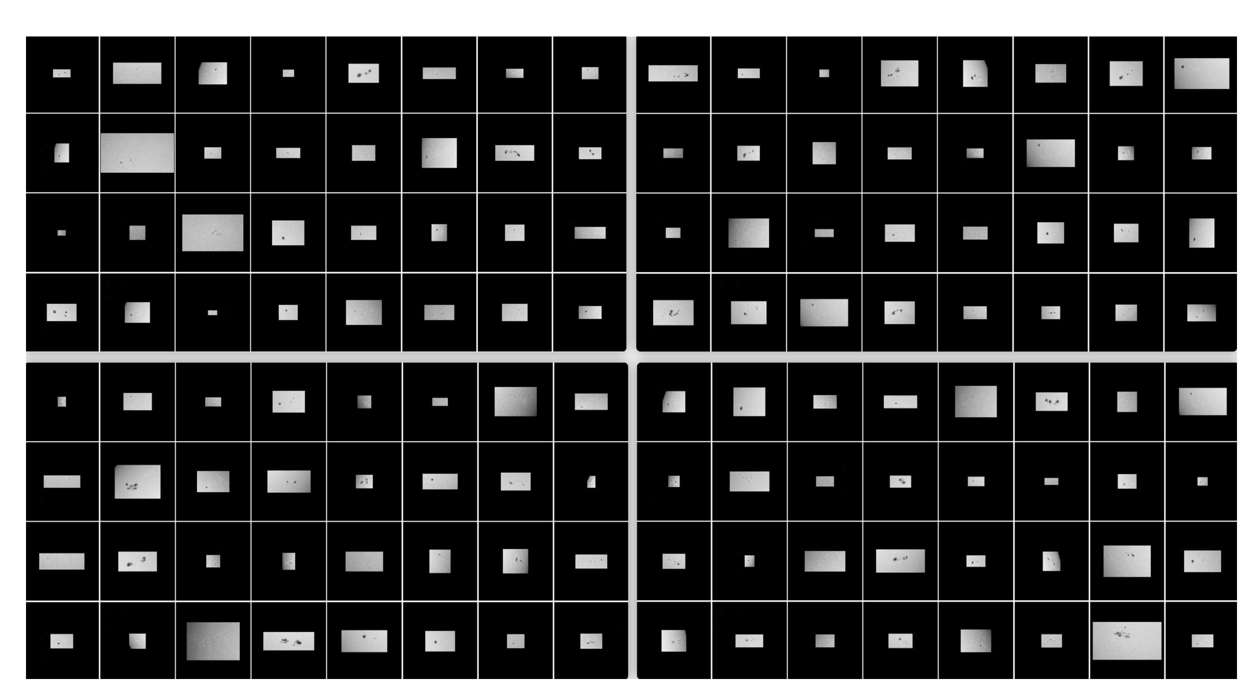


图 6　newSharp 中太阳黑子图像数据视场统一化示例

Fig. 6　FOV unification example of sunspot image data from newSharp

3 实验与对比分析

3.1 数据分布对比

将扩增并清洗后的 newSharp 与 Sharp 进行比较，图 7 为扩增前后数据分布情况的对比。其中 Sharp 一共 54 类，newSharp 在其基础上扩增了 3 类，共 57 类数据，扩增部分分别是 Cki，Fac 和 Fhi。其次，可以看出，newSharp 中大多数类别的数量较 Sharp 有所增加，例如 Axx 和 Cso 均增加超过 5 倍。此外，newSharp 增加了 Sharp 中少类的数量，例如 Fao 类的数量从 1 增加到 17，Cko 类的数量从 2 增加到 122，Eho 类的数量从 1 增加到 139，Cho 类的数量从 2 增加到 177，可见扩增对有效的少类样本数量增加有一定成效。然而，newSharp 中依然存在少数类别的数量较低，这是由于其本身在现实中较为罕见，例如 Ero 和 Chi。newSharp 中出现部分类别数量降低的原因可能是这部分数据在原 Sharp 中质量不高而被清洗过滤，例如 Fho 类的数量从 15 降低到 3，Fsc 类的数量从 11 降到 1。可以看到，即使扩增后部分类别数量增加甚至超过 5 倍的 newSharp，其数据依然呈现长尾分布，具有较为严重的类别不均衡，故本文与以往工作一样，将主要进行 McIntosh-Zpc 分类实验。

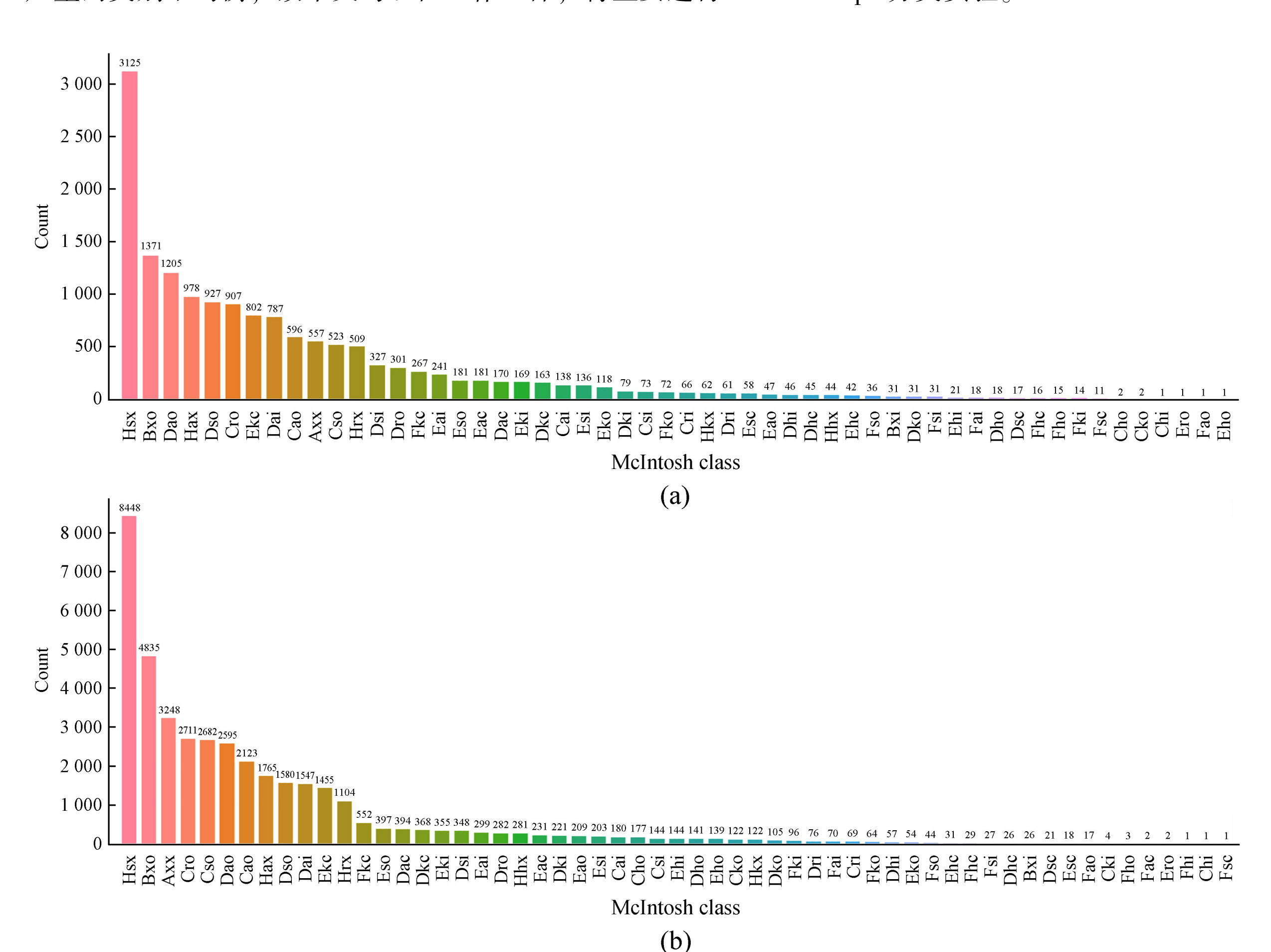


图 7 Sharp 和 newSharp 的数据分布对比。(a)Sharp 数据分布；(b)newSharp 数据分布

Fig. 7 Data distribution comparison between Sharp and newSharp. (a) Data distribution of Sharp; (b) data distribution of newSharp

另一方面，如图 8，是 Sharp 和 newSharp 的活动区分布情况对比，我们可以清晰观察到，newSharp 的活动区数量较 Sharp 增多，填补了 Sharp 大部分空缺，分布更加均匀，说明 newSharp 是较为完整可靠的。

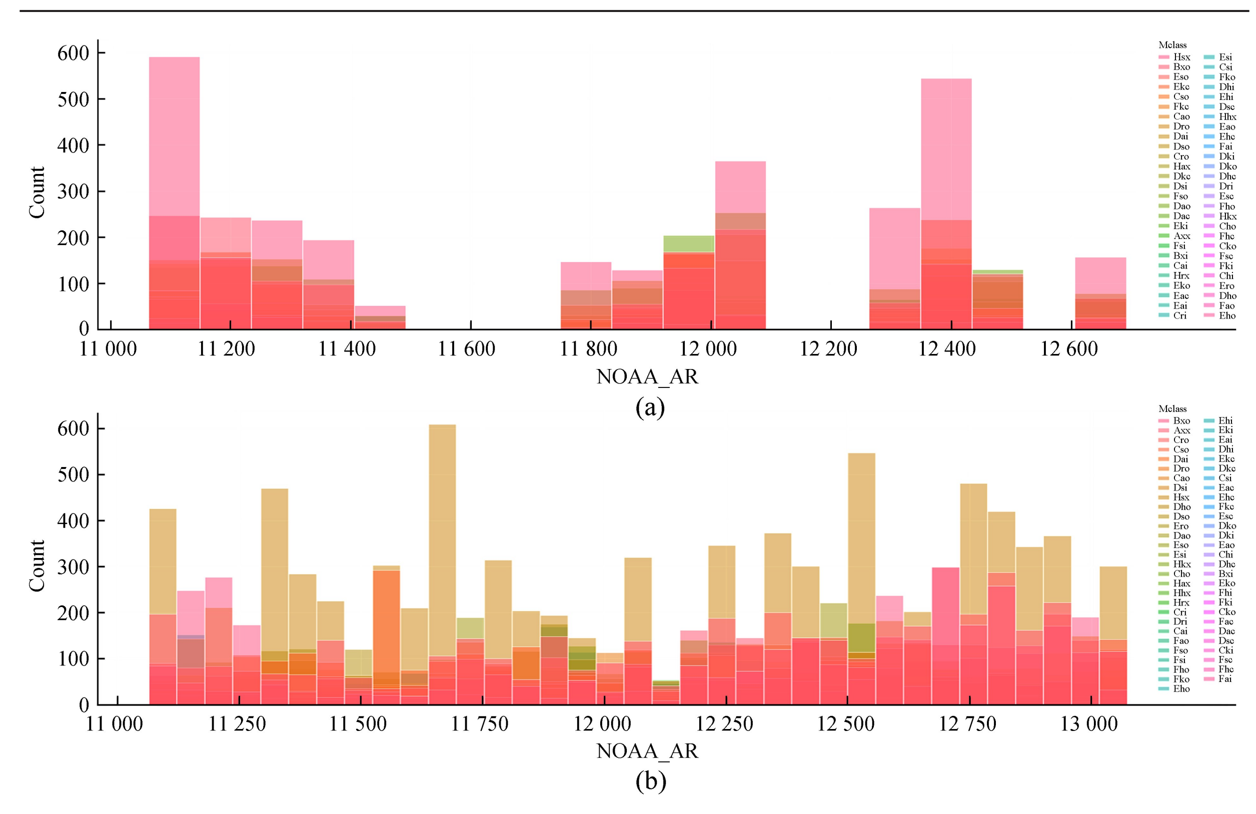


图 8　Sharp 和 newSharp 的活动区分布对比。(a)Sharp 活动区分布；(b)newSharp 活动区分布

Fig. 8　Active region distribution comparison between Sharp and newSharp. (a) Active region distribution of Sharp; (b) active region distribution of newSharp

对于 McIntosh-Zpc 数据分布而言，以 McIntosh-Z 为例，其中 Sharp 数据分布如图 9(a)，在 Sharp 基础上进行扩充后的原始 newSharp 数据分布如图 9(b)。由图 9 可以发现，原始 newSharp 中每一个类别的数量较 Sharp 均有明显增加，例如少类 F 类数量增加了超过 2. 8 倍，A 类数量甚至增加到原来的 17 倍；而经过一系列数据清洗之后的 newSharp 分布情况如图 9(c)，每一个类别的数量较 Sharp 相比也是只多不少，例如少类 F 类数量增加超过 1. 8 倍，而 A 类数量依然增加超过 5 倍。同样 newSharp 相比 Sharp 的数量增加也体现在 McIntosh-p 中的少类 h 类和 McIntosh-c 中难类 i 类上，如图 10 和图 11，分别增加了 4 倍和 1. 6 倍多，侧面印证了前期数据扩增和数据清洗的操作是有效的，具体在数据层面分类精度的提升效果需通过实验验证。

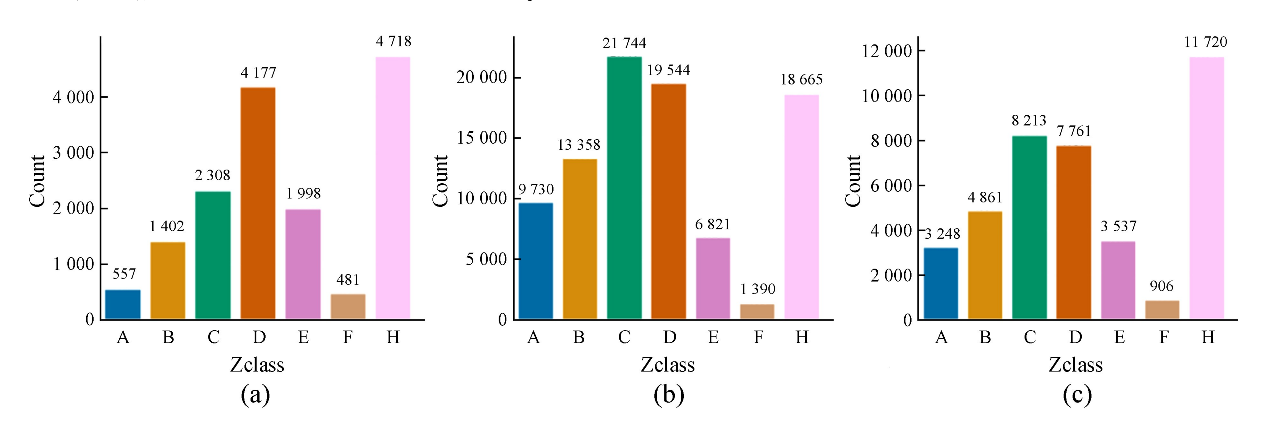


图 9　Sharp 和 newSharp 的 McIntosh-Z 分类数据分布对比。(a)Sharp 数据分布；(b)原始 newSharp 数据分布；(c)清洗后 newSharp 数据分布

Fig. 9　Data distribution comparison of McIntosh-Z classification between Sharp and newSharp. (a) Data distribution of Sharp; (b) data distribution of original newSharp; (c) data distribution of cleaned newSharp

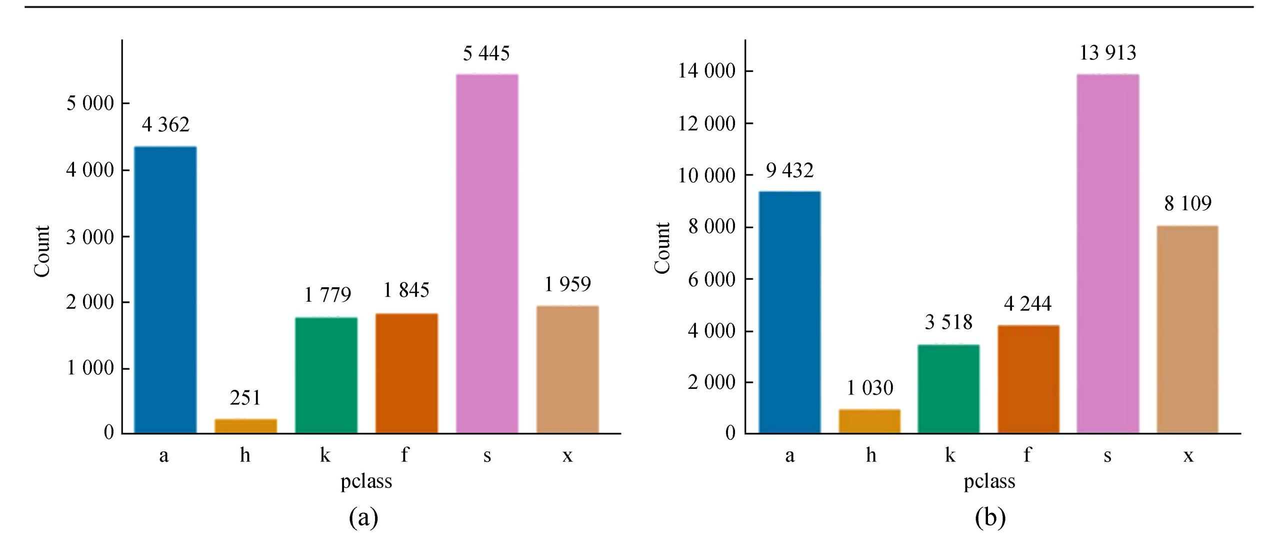


图 10　Sharp 和 newSharp 的 McIntosh-p 分类数据分布对比。(a)Sharp 数据分布；(b)清洗后 newSharp 数据分布

Fig. 10　Data distribution comparison of McIntosh-p classification between Sharp and newSharp. (a) Data distribution of Sharp; (b) data distribution of cleaned newSharp

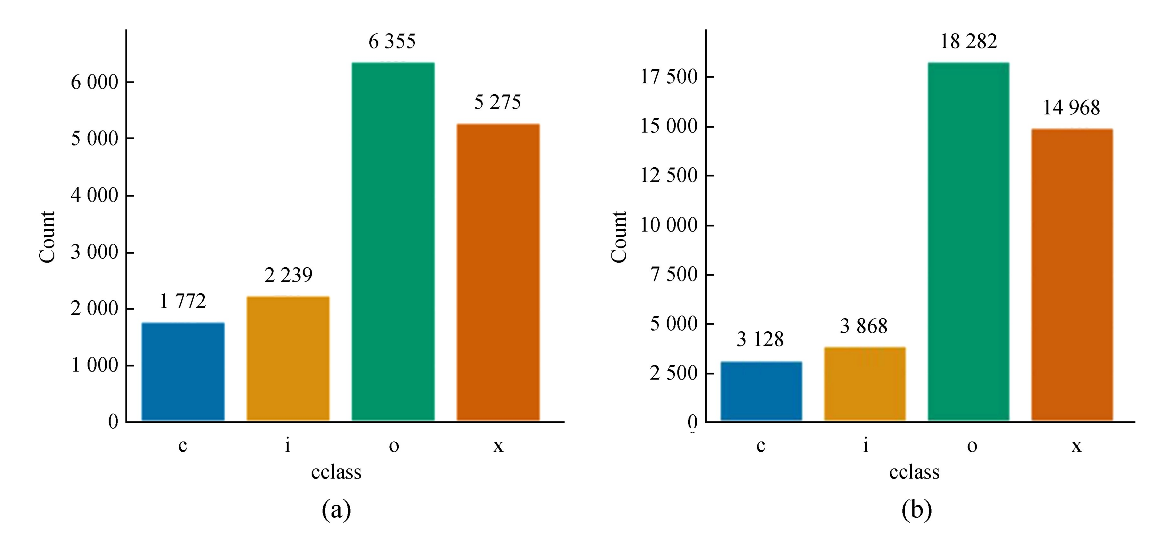


图 11　Sharp 和 newSharp 的 McIntosh-c 分类数据分布对比。(a)Sharp 数据分布；(b)清洗后 newSharp 数据分布

Fig. 11　Data distribution comparison of McIntosh-c classification between Sharp and newSharp. (a) Data distribution of Sharp; (b) data distribution of cleaned newSharp

3.2　基于 Sharp 的数据分配方式对比与选择

以往大多数太阳黑子的麦金托什分类工作中选择简单地将数据随机划分为训练集、验证集、测试集进行实验。但是由于太阳黑子演变的连续性，同时在数据清洗阶段发现不少太阳黑子在 SRS 信息更新之前已经完成演化，最终这部分数据因为麦金托什标签信息的准确性存疑而舍去。连续演化的黑子图像之间存在较大相似性，故简单随机划分数据的方式有可能造成数据集彼此交叉污染。经过观察，按照太阳黑子所在的活动区编号进行划分可使得数据集之间相互独立。

本文基于 Sharp 数据库并使用 4 种经典的分类网络模型(LeNet-5，AlexNet，VGG16 和 ResNet-18)针对两种数据分配方式进行 McIntosh-Zpc 分类对比实验：随机划分数据集(Random)和按活动区划分数据集(AR)，分别按两种分配方式将 Sharp 数据库划分为 70%训练集、20%验证集和 10%测试集。两种方式划分 McIntosh-Z 实验的具体数据分布如表 3 和表 4，两种方式划分 McIntosh-p 实验的具体数据分布如表 5 和表 6，两种方式划分 McIntosh-c 实验的具体数据分布如表 7 和表 8。

表 3　基于 Sharp 和按活动区分配的 McIntosh-Z 数据分布

Table 3　McIntosh-Z data distribution based on Sharp and AR-partition

	A	B	C	D	E	F	H	Total
Train-set	390	982	1 665	2 967	1 474	218	3 346	11 042
Val-set	92	333	489	926	360	63	963	3 226
Test-set	75	87	149	284	162	17	409	1 183
Total	557	1 402	2 303	4 177	1 996	298	4 718	15 451

表 4　基于 Sharp 和随机分配的 McIntosh-Z 数据分布

Table 4　McIntosh-Z data distribution based on Sharp and Random-partition

	A	B	C	D	E	F	H	Total
Train-set	407	1 015	1 634	3 005	1 423	214	3 344	11 042
Val-set	102	292	480	852	418	64	1 018	3 226
Test-set	48	95	189	320	155	20	356	1 183
Total	557	1 402	2 303	4 177	1 996	298	4 718	15 451

表 5　基于 Sharp 和按活动区分配的 McIntosh-p 数据分布

Table 5　McIntosh-p data distribution based on Sharp and AR-partition

	a	h	k	r	s	x	Total
Train-set	3 094	162	1 246	1 328	3 840	1 372	11 042
Val-set	966	33	286	414	1 102	425	3 226
Test-set	283	21	159	102	456	162	1 183
Total	4 343	216	1 691	1 844	5 398	1 959	15 451

表 6　基于 Sharp 和随机分配的 McIntosh-p 数据分布

Table 6　McIntosh-p data distribution based on Sharp and Random-partition

	a	h	k	r	s	x	Total
Train-set	3 094	165	1 201	1 348	3 839	1 395	11 042
Val-set	935	37	355	361	1 127	411	3 226
Test-set	314	14	135	135	432	153	1 183
Total	4 343	216	1 691	1 844	5 398	1 959	15 451

表 7　基于 Sharp 和按活动区分配的 McIntosh-c 数据分布

Table 7　McIntosh-c data distribution based on Sharp and AR-partition

	c	i	o	x	Total
Train-set	1 258	1 617	4 431	3 736	11 042
Val-set	310	464	1 397	1 055	3 226
Test-set	177	125	397	484	1 183
Total	1 745	2 206	6 225	5 275	15 451

表 8　基于 Sharp 和随机分配的 McIntosh-c 数据分布

Table 8　McIntosh-c data distribution based on Sharp and Random-partition

	c	i	o	x	Total
Train-set	1 257	1 577	4 430	3 778	11 042
Val-set	351	470	1 303	1 102	3 226
Test-set	137	159	492	395	1 183
Total	1 745	2 206	6 225	5 275	15 451

实验中各参数统一设置为 epoch = 50，learning rate = 1e-4，batchsize = 16，评价指标选择 weighted-F_1 score。在预训练模型的过程中，为保证最终对比的合理性和有效性，本文使用 4 种模型分类结果的平均值作为两种数据分配方式的表现指标(均作四舍五入处理)，并且不额外进行数据增强和正则化操

作；同时为保证模型充分接触数据且避免最初训练误差过大，设置当 epoch = 10 时，开始保存验证精度最高的模型(后续实验同此操作)。最终的结果如图 12(a)，对于 McIntosh-Zpc 实验整体表现来看，Random 分配方式下的加权 F_1 分数平均可达 94%，而 AR 分配方式下的加权 F_1 分数平均仅有 49%，两者在 McIntosh-p 中甚至相差 52%。具体分析，如图 12(b)，(c)和(d)，分别是关于 McIntosh-Z/p/c 分类的实验结果，其中 McIntosh-Z 的 F 类由于数量少，在 AR 分配方式下的加权 F_1 分数为 0，而在 Random 分配方式下的加权 F_1 分数却高出 94%，足以体现 Random 方式由于太阳黑子图像的连续性使得结果显著虚高；同样 McIntosh-p 的 h 类在 AR 分配方式下的加权 F_1 分数也为 0，在 Random 分配方式下的加权 F_1 分数却高出 86%；此外，McIntosh-c 中的 i 属于难类，在 AR 分配方式下的加权 F_1 分数仅有 22%，但在 Random 分配方式下的加权 F_1 分数依然高于 72%。由此可见，简单地使用 Random 的分配方式来划分太阳黑子数据必然会掩盖其中关于少类难类的诸多问题，因此本文采取 AR 的分配方式进行实验。

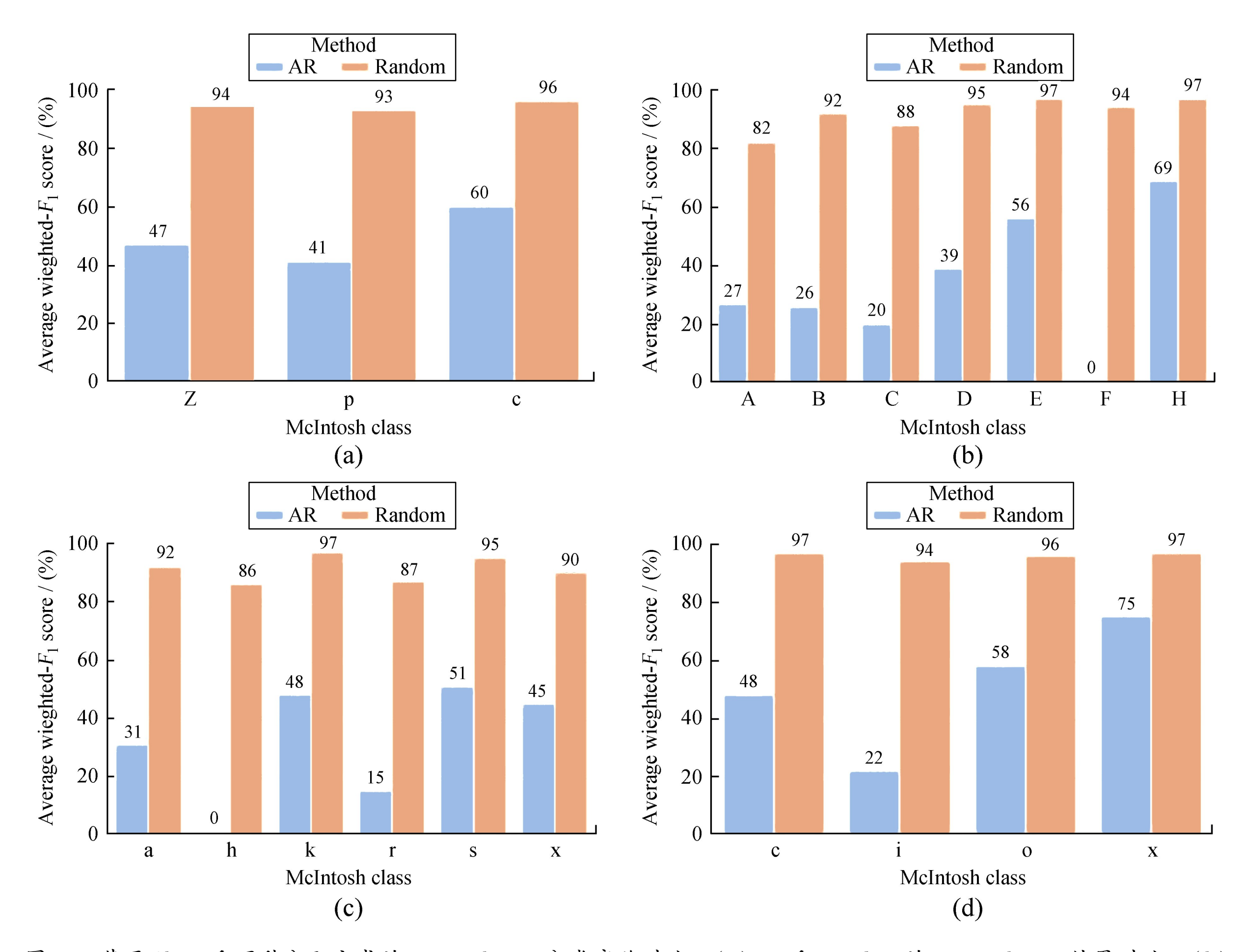


图 12 基于 Sharp 和两种分配方式的 McIntosh-Zpc 分类实验对比。(a) AR 和 Random 的 McIntosh-Zpc 结果对比；(b) AR 和 Random 的 McIntosh-Z 结果对比；(c) AR 和 Random 的 McIntosh-p 结果对比；(d) AR 和 Random 的 McIntosh-c 结果对比

Fig. 12 Result comparison of McIntosh-Zpc classification between AR-partition and Random-partition based on Sharp. (a) Result comparison of McIntosh-Zpc between AR and Random; (b) result comparison of McIntosh-Z between AR and Random; (c) result comparison of McIntosh-p between AR and Random; (d) result comparison of McIntosh-c between AR and Random

3.3 基于 newSharp 的 McIntosh-Zpc 分类实验对比

首先，按活动区分配方式合理划分 newSharp 数据集。同 3.2 中活动区划分 Sharp 的方式，对于 newSharp 中每一个麦金托什类别，按照活动区分开，即每个类中计算各活动区数量，并从高到低进行排序(活动区非连续)，其中应舍弃活动区数量少于 3 的类别，最后统计有 7 个，分别是 Ero，Fho，Chi，Fhi，Fac，Fsc 和 Fhc。其次，由于扩增了一定幅度的图像数据，且为了网络模型学习到更多样本

特征，以 7∶2∶1 的方式划分数据集，将各个类别按活动区数量从高到低、由前往后依次选取 70%训练数据、20%验证数据、10%测试数据，即满足训练集数据量后再划分验证集，满足验证集数据量后将剩余数据划分为测试集。由此，经过基于活动区划分数据集后，得到有 40 207 张太阳黑子图像数据的 newSharp 数据库，具体 McIntosh-Z/p/c 分类实验数据分布情况如表 9、表 10 和表 11。

表 9　基于 newSharp 和按活动区分配的 McIntosh-Z 数据分布

Table 9　McIntosh-Z data distribution based on newSharp and AR-partition

	A	B	C	D	E	F	H	Total
Train-set	2 283	3 407	5 818	5 502	2 554	678	8 299	28 541
Val-set	600	1 068	1 901	1 785	714	142	2 634	8 844
Test-set	365	386	493	474	267	50	787	2 822
Total	3 248	4 861	8 212	7 761	3 535	870	11 720	40 207

表 10　基于 newSharp 和按活动区分配的 McIntosh-p 数据分布

Table 10　McIntosh-p data distribution based on newSharp and AR-partition

	a	h	k	r	s	x	Total
Train-set	6 687	742	2 578	3 007	9 837	5 690	28 541
Val-set	2 232	208	667	949	3 120	1 668	8 844
Test-set	511	46	273	286	955	751	2 822
Total	9 430	996	3 518	4 242	13 912	8 109	40 207

与 3.2 中实验操作相同，使用 LeNet-5，Alex Net，VGG16 和 ResNet-18 四种经典的分类网络模型对 newSharp 进行 McIntosh-Zpc 分类实验，将分类结果取平均值作为 newSharp 的表现指标(均作四舍五入处理)，并和按活动区分配的 Sharp 结果进行对比来验证太阳黑子数据扩充和清洗等操作的有效性。为保证对比的合理性，本实验不进行额外的数据增强和模型正则化，同时各项参数设置一致：epoch = 50，learning rate = 1e-4，batchsize = 16，评价指标选择 weighted-F_1 score。最终将 newSharp 分类结果与 Sharp 进行比较，如图 13。

表 11　基于 newSharp 和按活动区分配的 McIntosh-c 数据分布

Table 11　McIntosh-c data distribution based on newSharp and AR-partition

	c	i	o	x	Total
Train-set	2 223	2 820	12 916	10 582	28 541
Val-set	603	814	4 193	3 234	8 844
Test-set	270	232	1 168	1 152	2 822
Total	3 096	3 866	18 277	14 968	40 207

整体分析而言，如图 13(a)，newSharp 在 McIntosh-Z/c 中表现和 Sharp 相差不大，而在以往工作和 Sharp 中表现欠佳的 McIntosh-p 分类却表现较好，其加权 F_1 分数提高了 13%。具体分析，如图 13(b)，尽管 newSharp 在 McIntosh-Z 中对于常见的多类 H 类由于扩增数据引入更多复杂的黑子特征而导致表现与 Sharp 相比有所下降，但对于 A 类、B 类、C 类、D 类的加权 F_1 分数分别提高了 12%，5%，10%和 11%，甚至现实中较为罕见的少类 F 类的加权 F_1 分数从 0%提高到 5%，侧面体现出数据扩充后增加有用样本的一定成效；同样如图 13(d)，尽管 newSharp 在 McIntosh-c 的多类 x 类中由于数据扩充引入更多复杂特征使得表现较 Sharp 有所下降，但在其他类中均有提高，其中难类 i 类的加权 F_1 分数提高了 4%，c 类的加权 F_1 分数提高了 14%，侧面体现出扩增了这些类别中包含更多有用特征的数据样本；而在以往工作中表现较差的 McIntosh-p 分类中，newSharp 的表现相对较为理想，如图 13(c)，多类 a 类和 s 类的加权 F_1 分数分别比 Sharp 提高了 7%和 5%，对于 r 类、x 类、k 类的加权 F_1 分数分别提高 6%，21%和 22%，甚至少类 h 类的加权 F_1 分数从 0%突破到了 17%，极大体现出本次对于太阳黑子数据进行扩充和清洗等一系列操作的有效性与必要性。

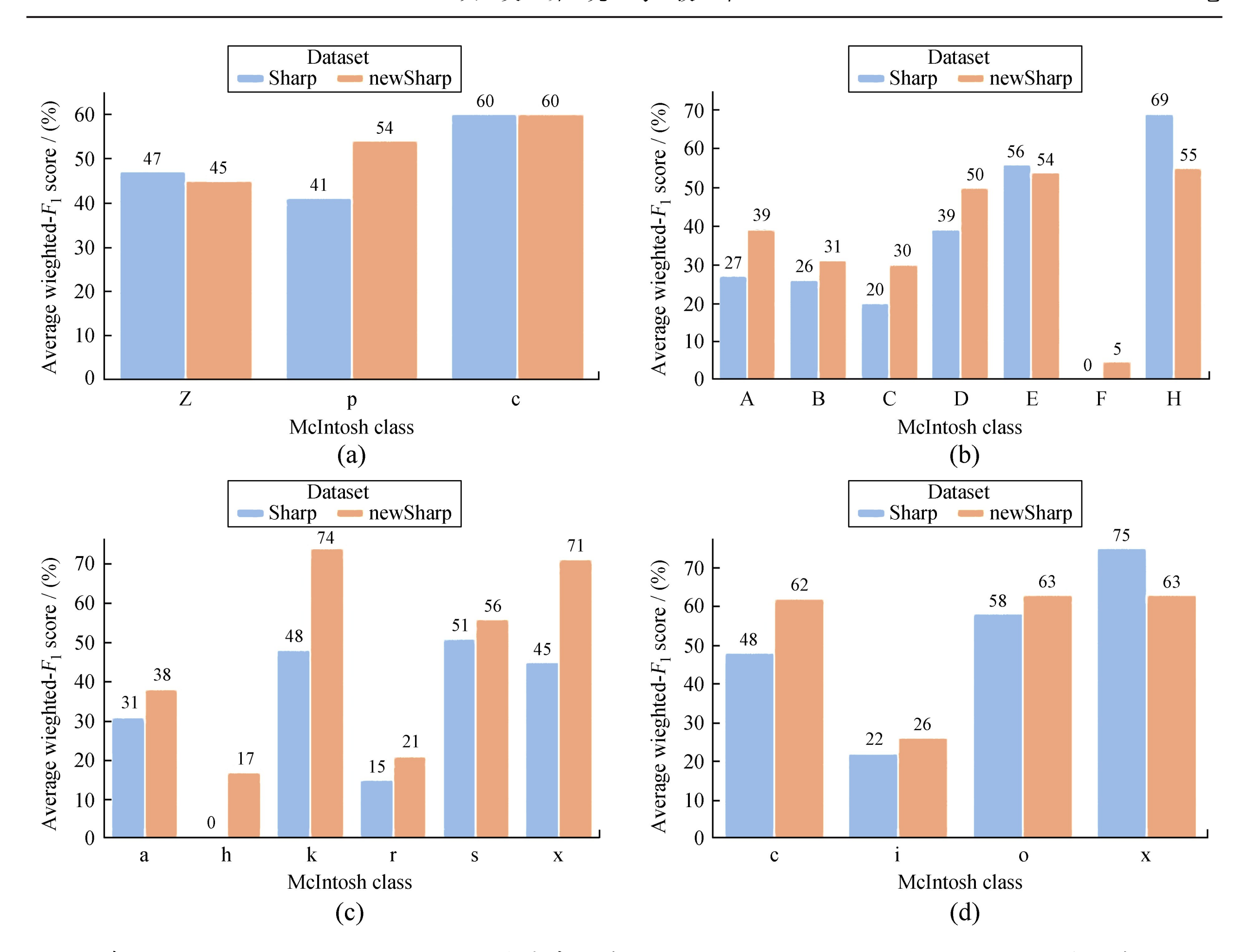


图 13 基于 Sharp 和 newSharp 的 McIntosh-Zpc 分类实验对比。(a) Sharp 和 newSharp 的 McIntosh-Zpc 结果对比；(b) Sharp 和 newSharp 的 McIntosh-Z 结果对比；(c) Sharp 和 newSharp 的 McIntosh-p 结果对比；(d) Sharp 和 newSharp 的 McIntosh-c 结果对比

Fig. 13 Result comparison of McIntosh-Zpc classification between Sharp and newSharp. (a) Result comparison of McIntosh-Zpc between Sharp and newSharp; (b) result comparison of McIntosh-Z between Sharp and newSharp; (c) result comparison of McIntosh-p between Sharp and newSharp; (d) result comparison of McIntosh-c between Sharp and newSharp

4 结 论

本文根据以往采用深度学习方法进行太阳黑子麦金托什分类时出现的问题与挑战，主要从数据和方法方面，建立一个完整太阳周期，且经过数据清洗，同时保留一定现实数据特征的太阳黑子数据库，以及使用一系列针对太阳黑子图像科学合理的实验预处理操作。最后通过在经典分类网络模型上进行测试实验，验证了数据库和实验方法的有效性，为后续使用深度学习实现基于复杂数据集且端到端的自动化太阳黑子麦金托什分类任务奠定坚实基础。

致谢：感谢太阳动力学天文台提供观测数据。

参考文献：

[1] 付小娜. 基于深度学习的太阳黑子群分类方法研究 [D]. 昆明：昆明理工大学, 2019.
FU X N. Research on sunspot classification method based on deep learning [D]. Kunming: Kunming University of Science and Technology, 2019.

[2] VALTONEN E. Space weather effects on technology [M]. Berlin: Springer. 2005: 241-273.

[3] BLANTER E, MOUËL J L L, PERRIER F, et al. Short-term correlation of solar activity and sunspot: evidence of lifetime increase [J]. Solar Physics, 2006, 237(2): 329-350.

[4] USOSKIN I G, KOVALTSOV G A, CHATZISTERGOS T. Dependence of the sunspot-group size on the level of solar activity and its influence on the calibration of solar observers [J]. Solar Physics, 2016, 291(12): 3793-3805.

[5] BERLYAND B. On the lifetime of sunspot groups during the maximum epoch of an 11-year solar cycle [J]. Byulletin Solnechnye Dannye Akademie Nauk SSSR, 1982, 1982: 96-99.

[6] 王雅妮. 利用卷积神经网络自动归类太阳黑子 [D]. 哈尔滨：哈尔滨工业大学, 2020.
WANG Y N. Automatic classification of sunspots using convolutional neural networks [D]. Harbin: Harbin Institute of Technology, 2020.

[7] LEE K, MOON Y J, LEE J Y, et al. Solar flare occurrence rate and probability in terms of the sunspot classification supplemented with sunspot area and its changes [J]. Solar Physics, 2012, 281(2): 639-650.

[8] 李泠, 崔延美, 刘四清, 等. 太阳黑子自动识别与特征参量自动提取 [J]. 空间科学学报, 2020, 40(3): 315-322.
LI L, CUI Y M, LIU S Q, et al. Automatic detection of sunspots and extraction of sunspot characteristic parameters [J]. Chinese Journal of Space Science, 2020, 40(3): 315-322.

[9] 赵梓良, 刘家真, 胡真, 等. 一种层次化的太阳黑子快速自动识别方法 [J]. 光电工程, 2020, 47(7): 39-49.
ZHAO Z L, LIU J Z, HU Z, et al. A hierarchical method for quick and automatic recognition of sunspots [J]. Opto-Electronic Engineering, 2020, 47(7): 39-49.

[10] NGUYEN S H, NGUYEN T T, NGUYEN H S. Rough set approach to sunspot classification problem [C] // Proceedings of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. 2005: 263-272.

[11] NGUYEN T T, WILLIS C P, PADDON D J, et al. Learning sunspot classification [J]. Fundamenta Informaticae, 2006, 72(1/3): 295-309.

[12] COLAK T, QAHWAJI R. Automatic sunspot classification for real-time forecasting of solar activities [C] // Proceedings of the 3rd International Conference on Recent Advances in Space Technologies. 2007: 733-738.

[13] COLAK T, QAHWAJI R. Automated McIntosh-based classification of sunspot groups using MDI images [J]. Solar Physics, 2008, 248(2): 277-296.

[14] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[15] NEBAUER C. Evaluation of convolutional neural networks for visual recognition [J]. IEEE Transactions on Neural Networks, 1998, 9(4): 685-696.

[16] SCHMIDHUBER J. Deep learning in neural networks: an overview [J]. Neural networks, 2015, 61: 85-117.

[17] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning [M]. Cambridge: MIT press, 2016.

[18] HALE G E, ELLERMAN F, NICHOLSON S B, et al. The magnetic polarity of sun-spots [J]. The Astrophysical Journal, 1919, 49: 153.

[19] WALDMEIER M. Coronal radiation and ionospheric variations during the solar eclipse, July 9, 1945 [J]. Terrestrial Magnetism and Atmospheric Electricity, 1947, 52(3): 333-338.

[20] CORTIE A. On the types of sun-spot disturbances [J]. The Astrophysical Journal, 1901, 13: 260.

[21] MCINTOSH P S. The classification of sunspot groups [J]. Solar Physics, 1990, 125(2): 251-267.

[22] KRIVSKY L. Solar activity observations and predictions [J]. Earth-Science Reviews, 1974, 10(4): 352.

[23] BORNMANN P L, SHAW D. Flare rates and the McIntosh active-region classifications [J]. Solar Physics, 1994, 150(1): 127-146.

[24] FANG Y H, CUI Y M, AO X Z. Deep learning for automatic recognition of magnetic type in sunspot groups [J]. Advances in Astronomy, 2019, 2019(123): 1-10.

[25] MASON J P, HOEKSEMA J. Testing automated solar flare forecasting with 13 years of Michelson Doppler Imager magnetograms [J]. The Astrophysical Journal, 2010, 723(1): 634.

[26] SCHRIJVER C J. A characteristic magnetic field pattern associated with all major solar flares and its use in flare forecasting [J]. The Astrophysical Journal, 2007, 655(2): L117.

[27] BOBRA M G, SUN X, HOEKSEMA J T, et al. The Helioseismic and Magnetic Imager (HMI) vector magnetic field pipeline: SHARPs-space-weather HMI active region patches [J]. Solar Physics, 2014, 289(9): 3549-3578.

[28] 柯大荣，赵永恒. 一种图象传输系统及其 FITS 数据基本格式 [J]. 现代图书情报技术，1994, 10(2): 22-27.
KE D R, ZHAO Y H. An image transport sysetem and its fits basic format [J]. New Technology of Library and Information Service, 1994, 10(2): 22-27.

Sunspot Data Collection and Experimental Validation for McIntosh Classification

Zhou Meilin[1,2,3], Zhong Libo[2,3*]

(1. University of Chinese Academy of Sciences, Beijing 100049, China; 2. Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China, Email: zhonglibo@ioe.ac.cn; 3. Key Laboratory of Adaptive Optics, Chinese Academy of Sciences, Chengdu 610209, China)

Abstract: As an important basis for predicting solar activity, the McIntosh classification of sunspots is used by more and more international institutions and astronomical institutes because some categories are closely related to flare eruption. With the rapid increase in the amount of data, automatic McIntosh classification of sunspots has become an urgent need. Using the 720s-SHARP series data products provided by SDO/HMI and SRS files from NOAA as images and labels for McIntosh classification, this paper first augmented valid samples of a complete solar cycle (time span of 12 years) and cleaned data to establish the sunspot database newSharp on the basis of the Sharp database with only 7-year data. Secondly, in view of the characteristics of sunspot images, a series of preprocessing operations such as data allocation by active region number were taken, and proved its rationality and necessity. Finally, four classical classification neural network models in CNN were used to compare Sharp and newSharp for McIntosh classification experiments. The results show that compared with Sharp, newSharp not only has a significant increase in the amount of data, but also has better weighted F_1 score of most categories by augmenting valid samples and cleaning invalid samples. Besides, the weighted F_1 score of categories with a small number from newSharp even has achieved a breakthrough of 0. Over all, the weighted F_1 score of McIntosh-p improved the most, which greatly verifies the effectiveness of establishing a complete and reliable database and proves the rationality of using scientific and reasonable experimental methods. Thus it is able to better automatedly realize the end-to-end McIntosh classification tasks of sunspot images that are actually observed.

Key words: sunspot; McIntosh classification; Convolutional Neural Network; Sharp dataset